M-DG Seminar: Multinomial Processing Tree Modeling

Basics of MPT Modeling

Summer semester 2020

Prof. Dr. Daniel Heck

M-DG: Multinomial Processing Tree Modeling

Part	Date	Topic	Literature
(A) Theory	Self study	A1) Introduction	Erdfelder et al. (2009)
		A2) Basics of MPT modeling	Batchelder & Riefer (1999)
		A3) The software multiTree	Moshagen (2010)
		A4) Hierarchical MPT modeling	Lee (2011) Heck et al. (2018)
(B) Application	15.5.*	B1) Questions & Practice with multiTree	Batchelder & Riefer (1986)
	20.5.*	B2) Workflow: Developing an MPT model	Jung et al. (2019)

* Web-Conference, 12:00 – 15:00, https://webconf.hrz.uni-marburg.de/b/dan-fvk-ha6

Basics of MPT Modeling

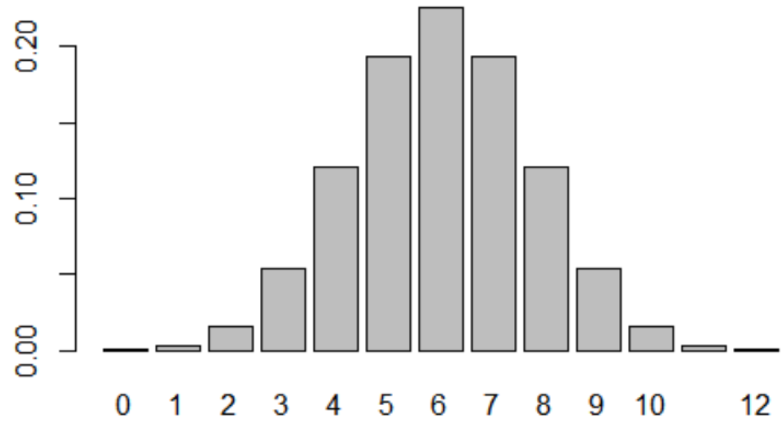
Overview:

1. Formal model structure
2. Identifiability
3. Parameter estimation
4. Model assessment & comparison
5. Appendix: The power divergence statistic

Binomial Models

Binomial model:

- One variable with 2 categories
- Data: Observed response frequencies $\boldsymbol{n} = (n_1, n_2)$
- Parameters: Vector of category probabilities $\boldsymbol{p} = (p_1, p_2)$
- Given independent sampling, the frequencies follow a binomial distribution:



Binomial Distribution

- N = 12 responses
- $p_1 = p_2 = 1/2$

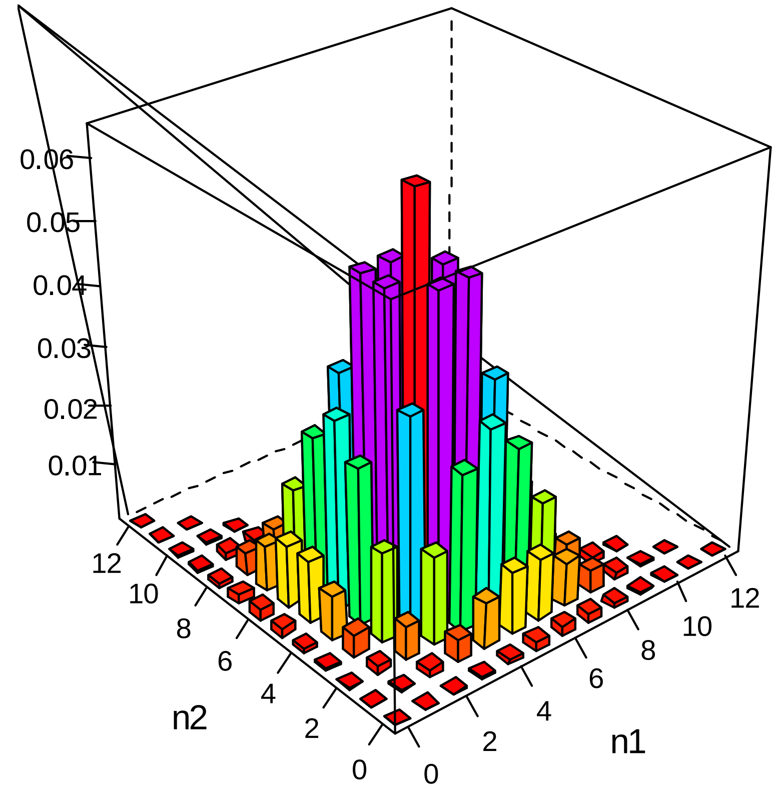
$$p(n_1, n_2) = \binom{N}{n_1} p_1^{n_1} (1 - p_1)^{N - n_1} = \frac{N!}{n_1! n_2!} p_1^{n_1} p_2^{n_2}$$

Multinomial Models

Multinomial model:

- One variable with *J* categories
- Data: Observed response frequencies $\boldsymbol{n} = (n_1, n_2, ..., n_J)$
- Parameters: Vector of category probabilities $\boldsymbol{p} = (p_1, p_2, ..., p_J)$
- Given independent sampling, the frequencies follow a multinomial distribution:

$$p(n_1, n_2, \ldots, n_J) = \frac{N!}{n_1! n_2! \cdots n_J!} p_1^{n_1} p_2^{n_2} \cdots p_J^{n_J}$$



Multinomial Distribution

- $N = 12$ responses
- $J = 3$ categories
- $p_1 = p_2 = p_3 = 1/3$

Parameterized Multinomial Models

Parameterized Multinomial Model:

- The category probabilities $\boldsymbol{p} = (p_1, p_2, ..., p_J)$ are rewritten as functions of the latent parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_S)$
- Based on the simple multinomial model, we define a set of model equations $f(\boldsymbol{\theta})$:
  - $p_1 = f_1(\theta_1, \theta_2, ..., \theta_S)$
  - $p_2 = f_2(\theta_1, \theta_2, ..., \theta_S)$
  - ....
  - $p_J = f_J(\theta_1, \theta_2, ..., \theta_S)$
- The set of possible values of S latent parameters θ_s is called parameter space Ω of the model.

Parameterized Multinomial Models

Example:

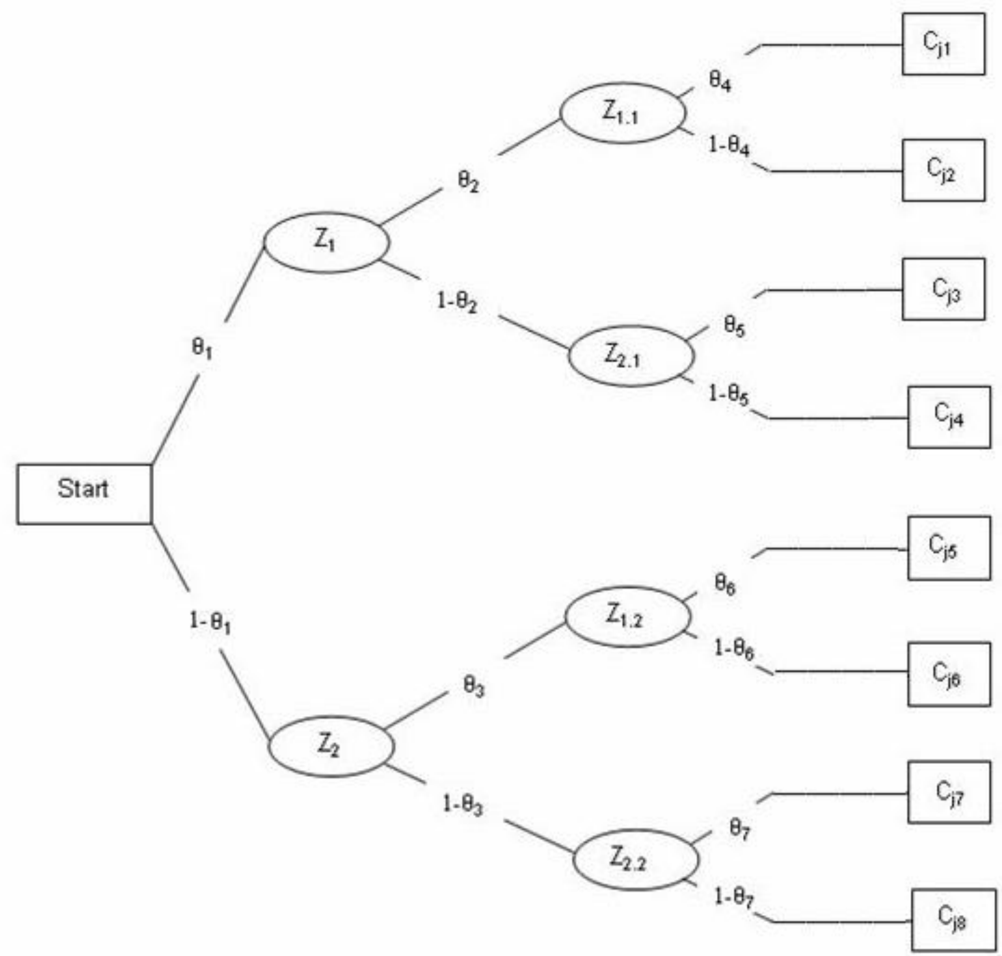
- Response categories in old-new recognition memory:
  1. *hit* = correct "old" response to old items
  2. *miss* = incorrect "new" response to old items
- Model equations of the 1-high threshold model (1HTM) of recognition memory

$$p(hit) = p(\text{"old"} \mid \text{old item}) = \boxed{r + (1-r)\,g}$$
$$p(miss) = p(\text{"new"} \mid \text{old item}) = \boxed{(1-r)\,(1-g)}$$

$$p(n_{hit}, n_{miss} \mid r, g) = \frac{N!}{n_{hit}!\, n_{miss}!} \boxed{[r + (1-r)g]}^{n_{hit}} \boxed{[(1-\mathrm{r})\,(1-\mathrm{g})]}^{n_{miss}}$$

MPT Models

- What distinguishes MPT models from other multinomial models?
- MPT models assume a specific form of the model equations
- Branch probabilities of a binary probability tree



Formal Definition of MPT Models

MPT models:

- A specific type of parameterized multinomial model
- Each parameter θ_s is in the interval [0, 1] (= a probability)
- The structure of the model equations is given as:

$$p_j = \sum_{i=1}^{I(j)} c_{ij} \prod_{s=1}^{S} \theta_s^{a_{ijs}} \cdot (1-\theta_s)^{b_{ijs}}, \qquad \sum_{j=1}^{J} p_j = 1, \qquad \theta_s \in [0, 1]$$

where s: Parameter index

j: Category index

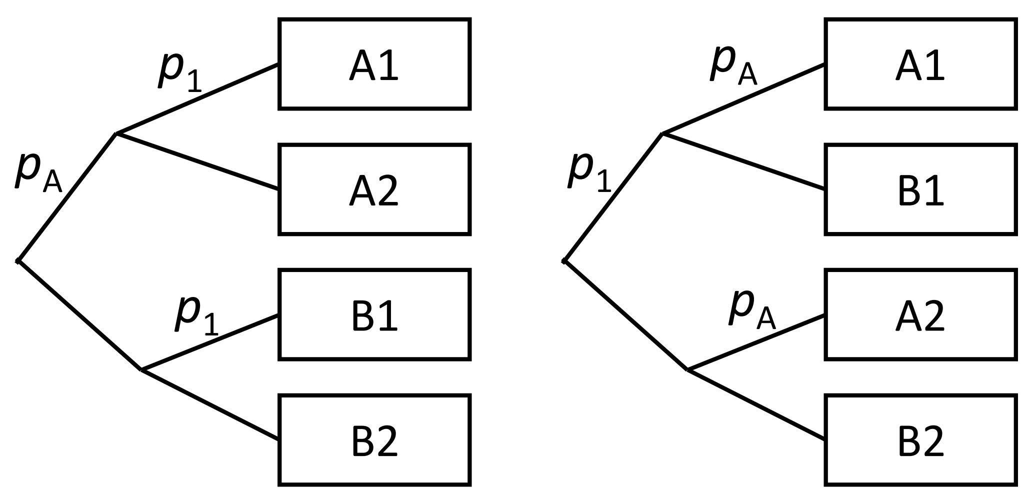
i: Branch index

c_{ij}: positive real number

a_{ijs}, b_{ijs}: nonnegative integer number (often 0 or 1)

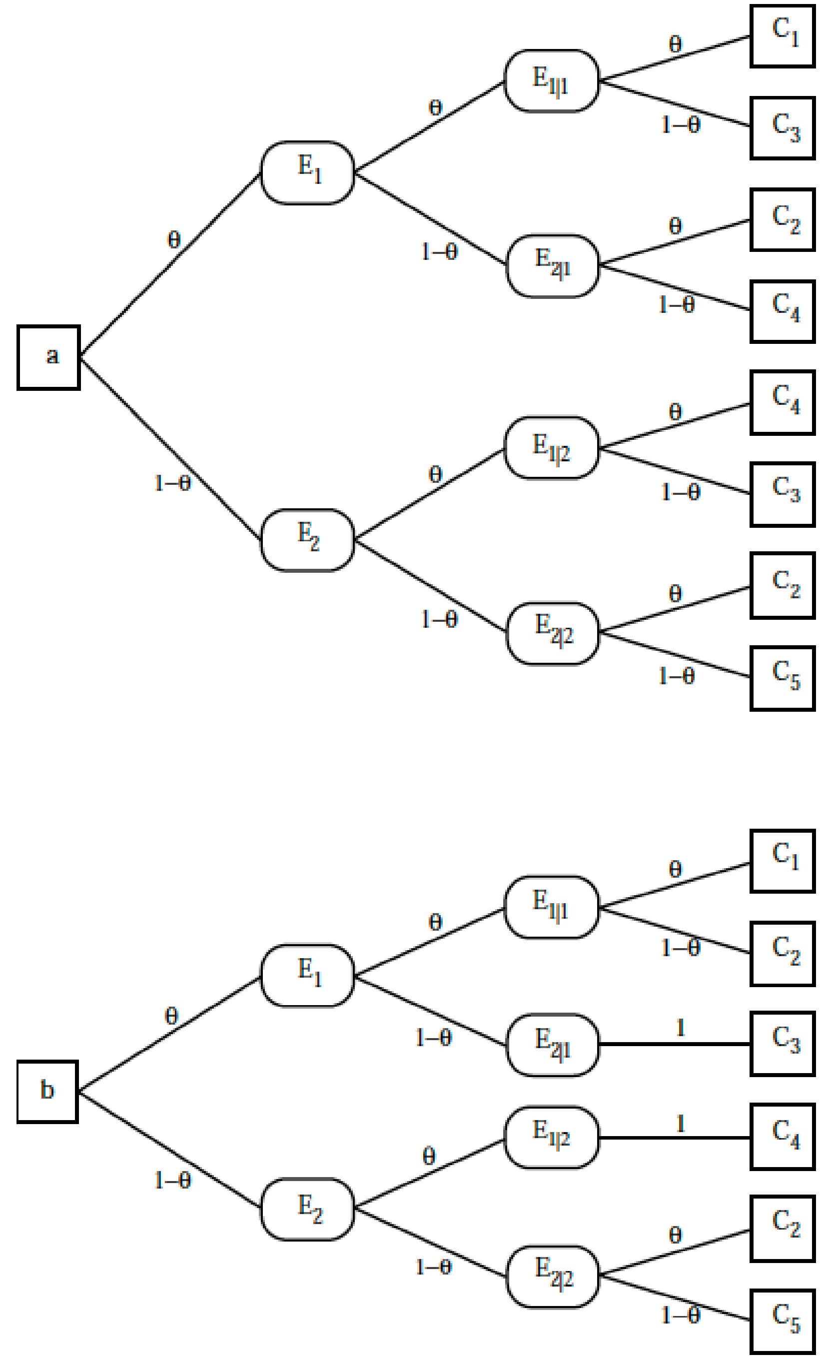
Uniqueness of the „Tree“

- A binary probabilistic event tree uniquely determines a system of MPT model equations.
- However: it is *not* true that any system of MPT model equations uniquely determines a specific tree diagram.
- Counter Example 1:
  Level switching in independence models
  - $p(\text{A1}) = p_\text{A}\, p_1$
  - $p(\text{A2}) = p_\text{A}\,(1 - p_1)$
  - $p(\text{B1}) = (1 - p_\text{A})\, p_1$
  - $p(\text{B2}) = (1 - p_\text{A})\,(1 - p_1)$



Counter Example 2

$$p_1(\theta) = \theta^3$$
$$p_2(\theta) = \theta \cdot (1 - \theta)$$
$$p_3(\theta) = \theta \cdot (1 - \theta)$$
$$p_4(\theta) = \theta \cdot (1 - \theta)$$
$$p_5(\theta) = (1 - \theta)^3$$



Basics of MPT Modeling

Overview:

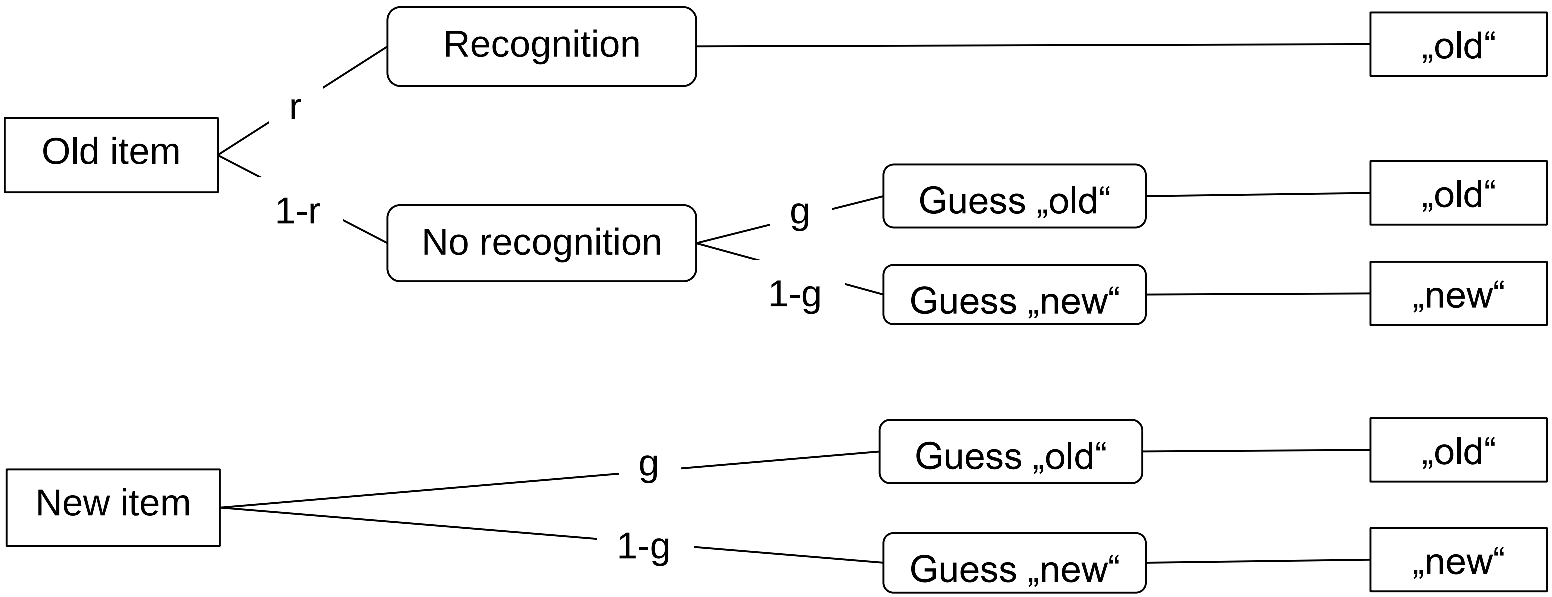
1. Formal model structure
2. Identifiability
3. Parameter estimation
4. Model assessment & comparison
5. Appendix: The power divergence statistic

Identifiability

- A MPT model defines a mapping $f: \Omega \rightarrow P$
  - Parameter space Ω = the set of all possible parameter vectors $\boldsymbol{\theta}$
  - Data space P (more precisely: space of category probabilities) = the set of all possible category probability vectors **p**

- Global identifiability:
  - A MPT model is globally identified if the mapping f is one-to-one for all parameters $\boldsymbol{\theta}$ in Ω.
- Local identifiability:
  - A MPT model is locally identified if the mapping f is one-to-one in the neighborhood of a specific point $\boldsymbol{\theta}_0$ in Ω.

1-High-Threshold Model (Blackwell, 1953)

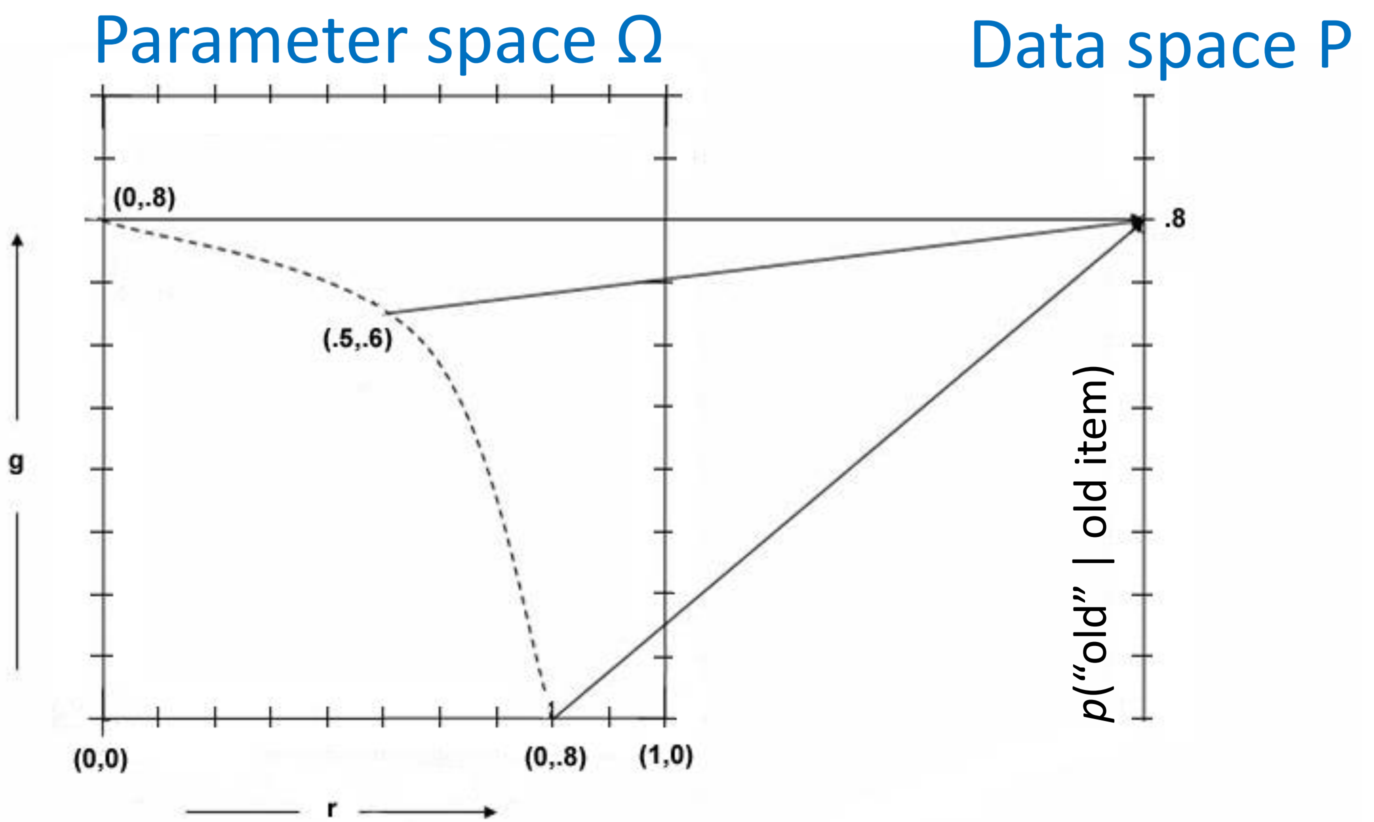


r = probability of recognition

g = probability of guessing „old“ given no recognition

Example 1: Nonidentifiability

- 1-high-threshold model limited to old items:
  $p(\text{“old”} \mid \text{old item}) = r + (1 - r)\, g$
- The model is *not* identified

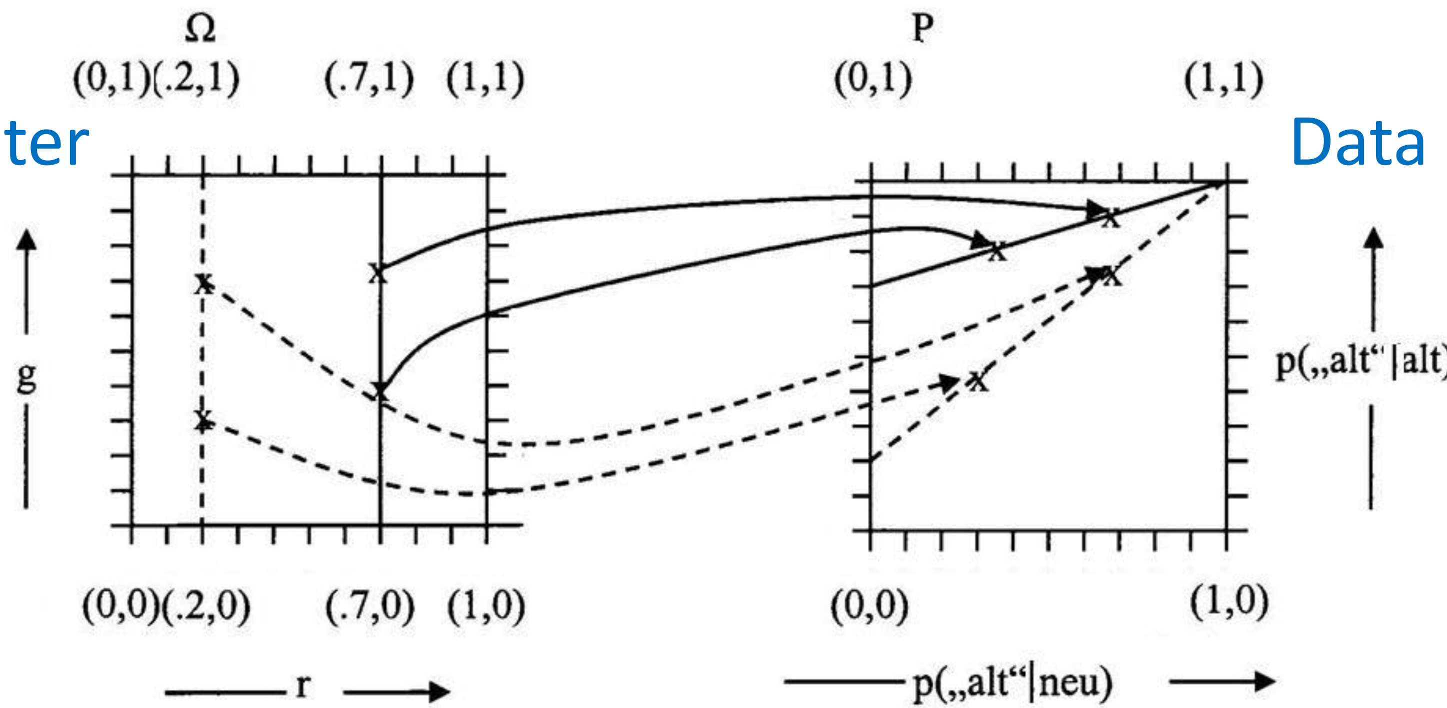


Example 2: Identifiability

- 1-high-threshold model for old & new items
  - p("old" | old item) = $r + (1 - r) \cdot g$
  - p(„old" | new item) = g
- The model *is* globally identified

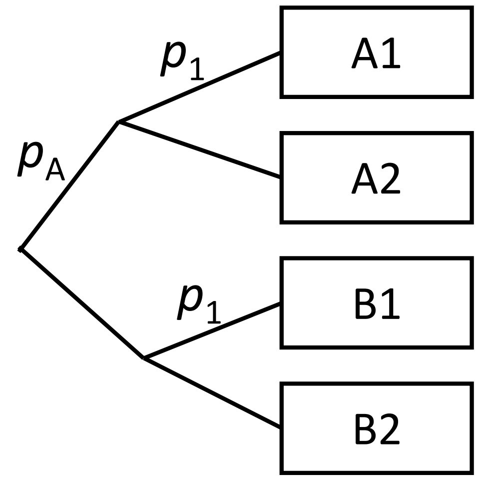
Parameter space

Data space



Identifiability: Two Important Theorems

Observable branches:



- A model is always globally identified if each of its branches terminates in a new empirical category (Hu & Batchelder, 1994).

Number of parameters:

- A model cannot have more parameters than degrees of freedom in the data.
- Hence, a necessary but not sufficient condition of identifiability for the number of parameters *S* is:

$$S \leq \sum_{k=1}^{K} (J_k - 1) \qquad [\, J_k = \text{number of response categories in condition } k]$$

Identifiability: Jacobian Matrix

Jacobian matrix

- Matrix of the first partial derivatives of all model equations with respect to all parameters θ_s
- r = maximum rank of the Jacobian matrix across Ω (can be computed in multiTree)

- General rules:
  - If $r < S$, then the model is neither locally nor globally identified.
  - If $r = S$, then the model is locally identified (but not necessarily globally).

Remedies for Nonidentifiable Models

How to get an identifiable MPT model?

a) Assume less parameters → Parameter constraints

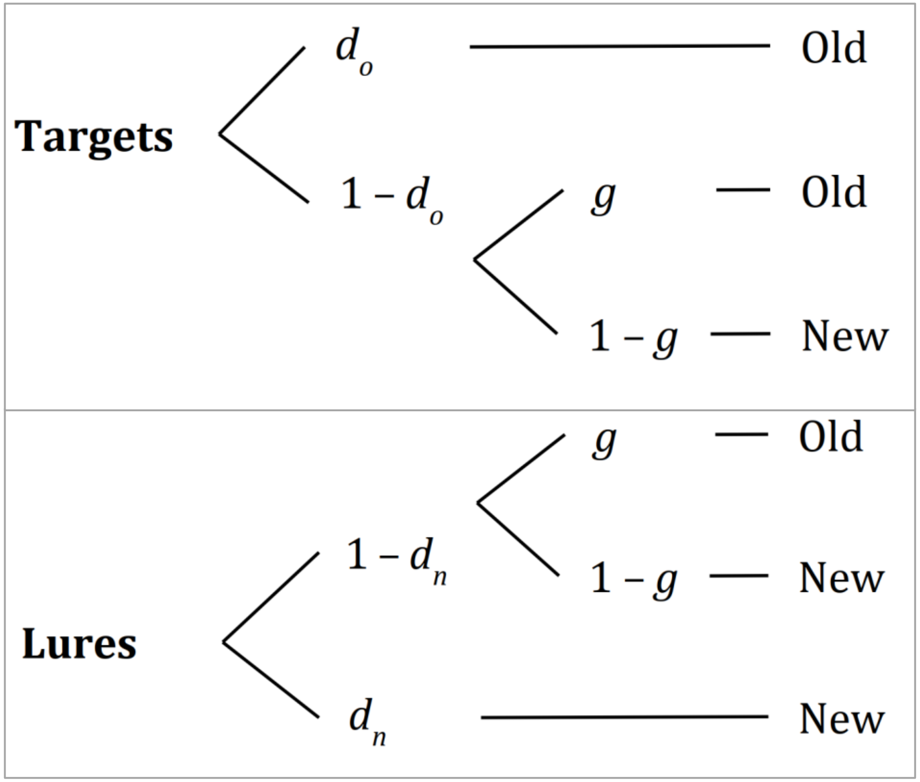
- Parameter fixations $\theta_s = c$ (with c = constant number)
- Equality constraints $\theta_s = \theta_t$ (for two parameters)

b) Add more empirical categories

- Additional conditions with no (or few) additional parameters
- Selective manipulations of parameters

Identifiability: Example

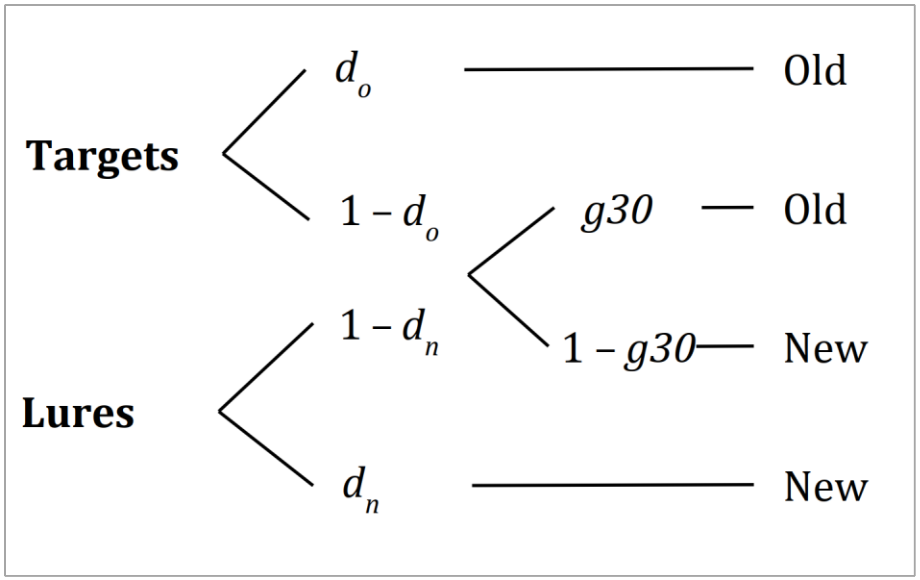
- 2-High Threshold Model
  - Parameters: $S = 3$ (d_o, d_n, g)
  - Free categories: $df = 2$
  - → not identifiable!



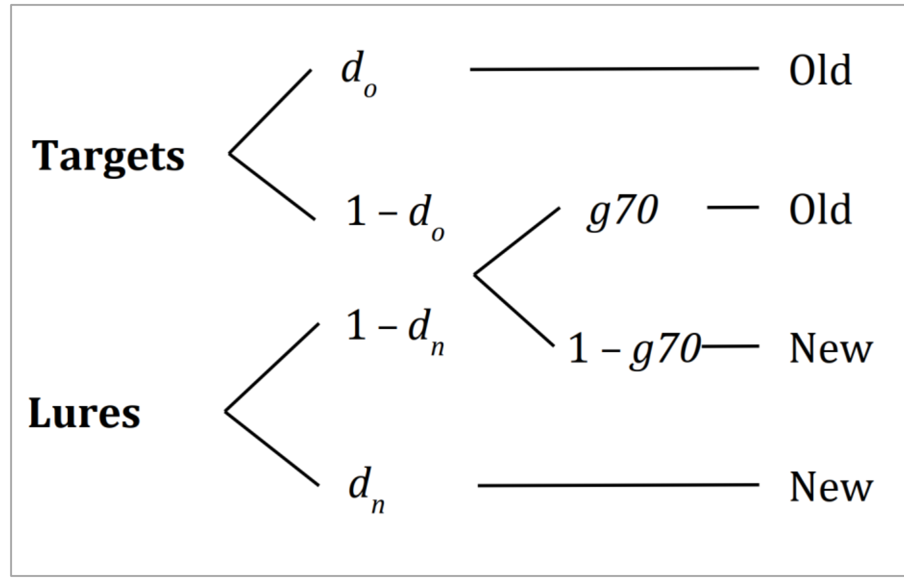
- Solutions:
  a) Constraints: Assume $d_o = d_n$
  b) More conditions: Base rate manipulation of response bias g

Identifiability: More Empirical Categories

(A) Base rate: 30% targets



(B) Base rate: 70% targets



→ *Two* additional degrees of freedom (*df* = 4)

→ But only *one* additional free parameter (*S* = 4)

→ Model is identifiable.

Basics of MPT Modeling

Overview:

1. Formal model structure
2. Identifiability
3. Parameter estimation
4. Model assessment & comparison
5. Appendix: The power divergence statistic

Parameter Estimation

- Given the data $n_1, \ldots, n_J$, what is the „best" vector of parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s, \ldots, \theta_S)$

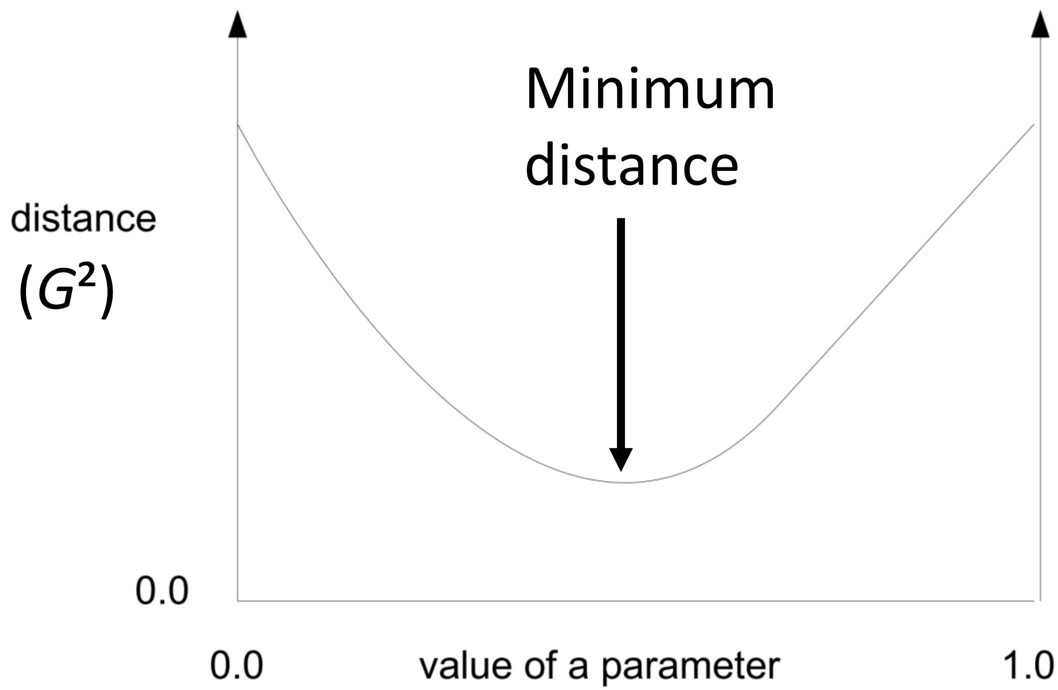
  → Find $\boldsymbol{\theta}$ that minimizes the distance between observed and expected category frequencies!

- Distance measure: The likelihood ratio statistic G^2

$$G^2(\theta) = -2 \sum_{j=1}^{J} n_j \ln\left(\frac{\overbrace{n_j}^{\text{observed frequencies}}}{\underbrace{N \cdot p_j(\theta)}_{\text{predicted frequencies}}}\right)$$

Parameter Estimation

- Which are the best parameter values given the data?
- Aim: Minimization of the distance measure G^2
- Example: MPT model with $S = 1$ parameter



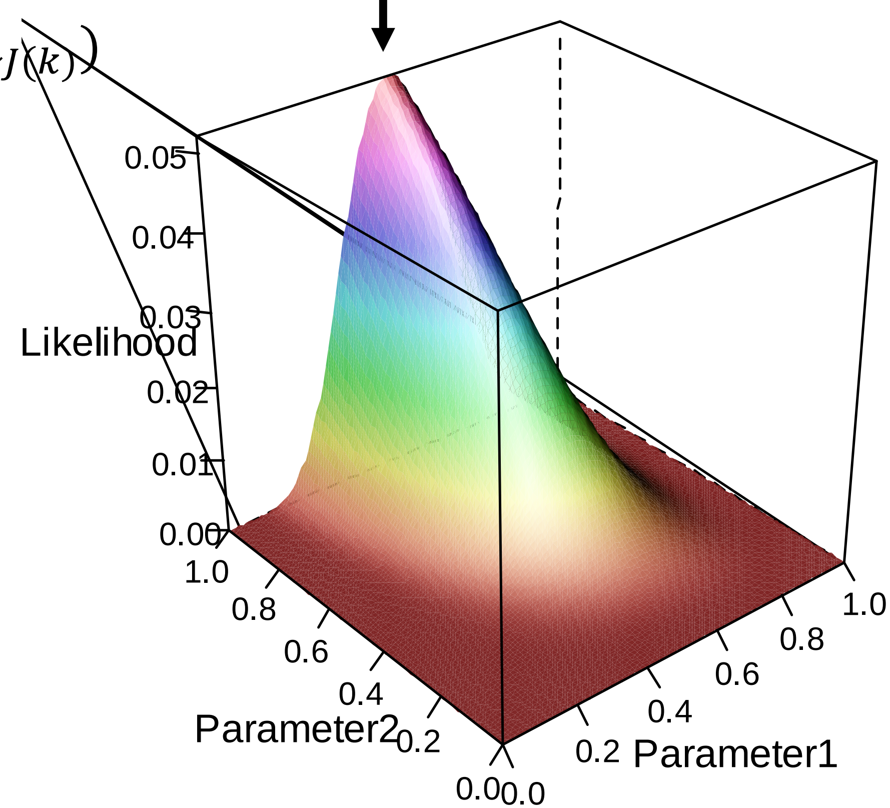
Parameter Estimation

- Minimization of G^2 is equivalent to maximizing the likelihood of the data given the parameters:

$$L(\theta\,; n) = \prod_{k=1}^{K} p_{N(k),\pi(k)}(n_{1(k)},\, n_{2(k)}\,, \ldots,\, n_{J(k)})$$

- Example: MPT model with S = 2 parameters

Maximum of the likelihood

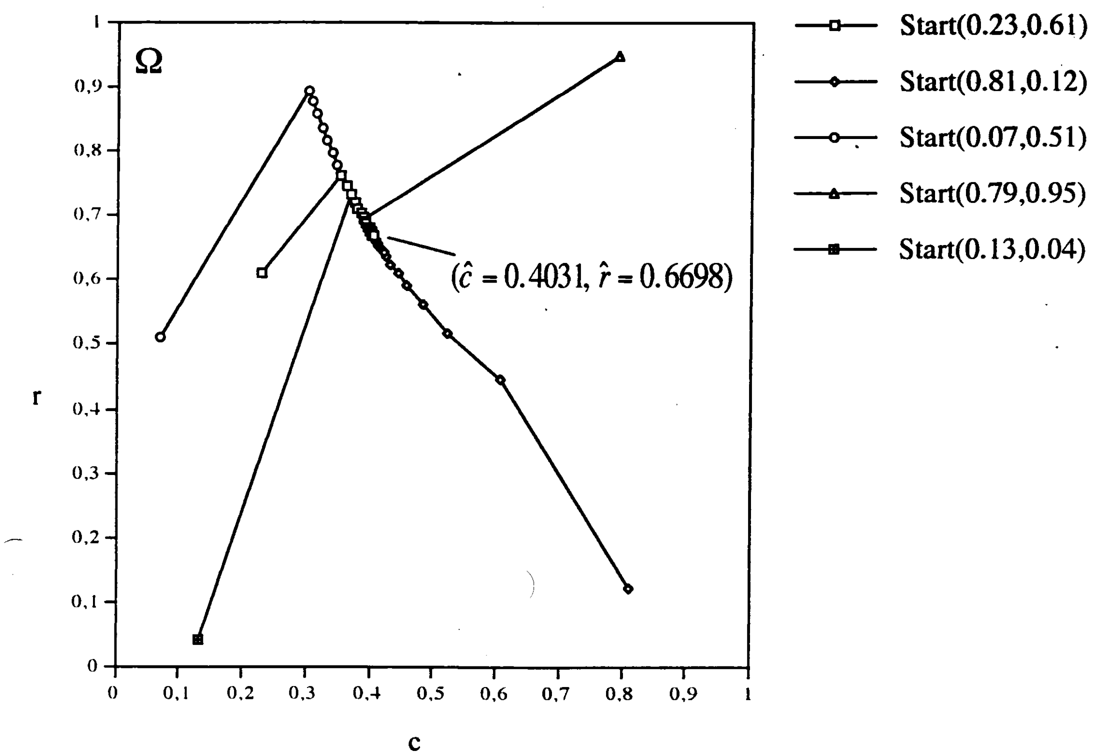


Expectation-Maximization-(EM) Algorithm

1. Choose a random start vector θ_i
2. **E**(xpectation)-Step:
   - Estimate the expected frequencies of the branches given θ_i and the observed category frequencies $n_{j(k)}$
3. **M**(aximization)-Step:
   - Let $i = i + 1$
   - Compute new G^2 estimates θ_i given the expected frequencies from step 2
4. Convergence?
   - if $|\theta_i - \theta_{i-1}| > \varepsilon$ → go back to Step 2 (ε = convergence criterion)
   - otherwise → accept θ_i as final parameter estimates

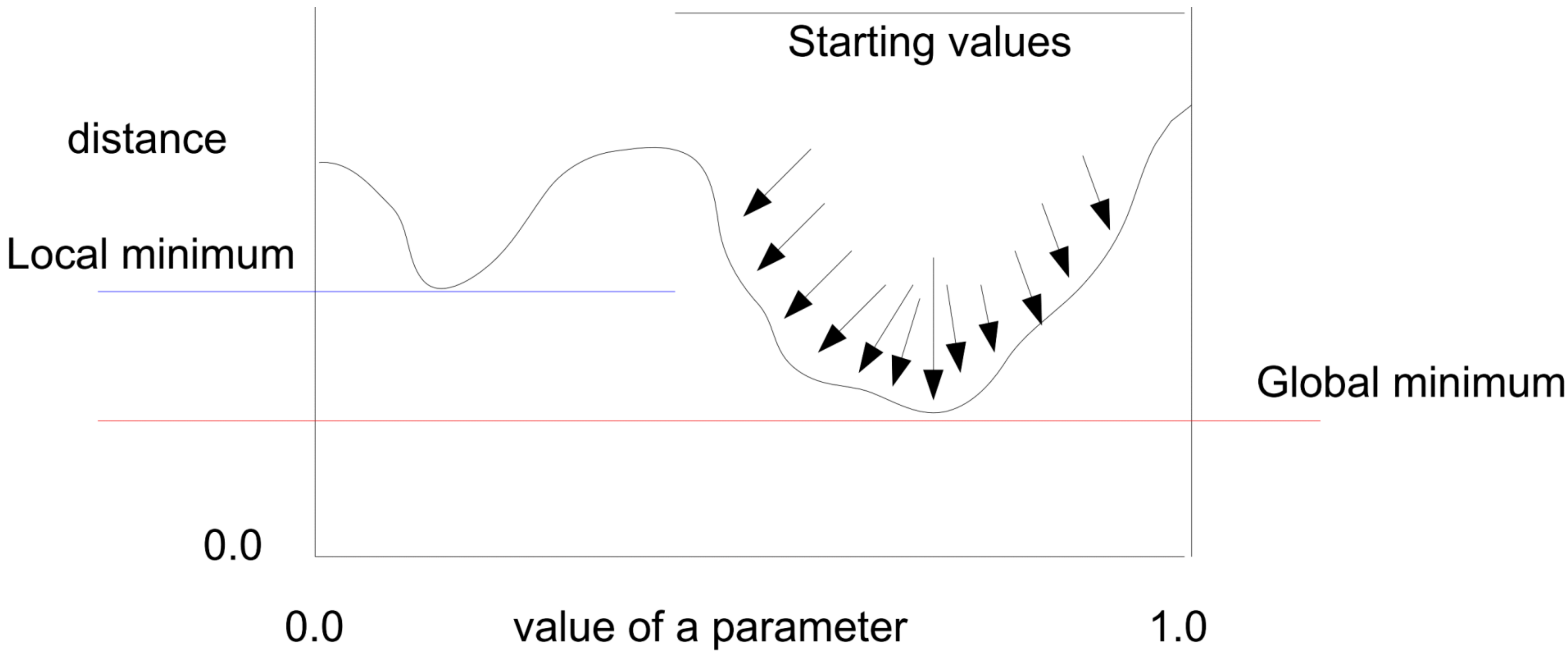
EM Algorithm

- Graphical illustration of the EM algorithm applied to an MPT model with S = 2 parameters
  - Different path of the EM algorithm for five starting values



Parameter Estimation: Local Minima

- Possible issue: Local minima of the likelihood function
- Solution: Fit model multiple times with random starting values



Basics of MPT Modeling

Overview:

1. Formal model structure
2. Identifiability
3. Parameter estimation
4. Model assessment & comparison
5. Appendix: The power divergence statistic

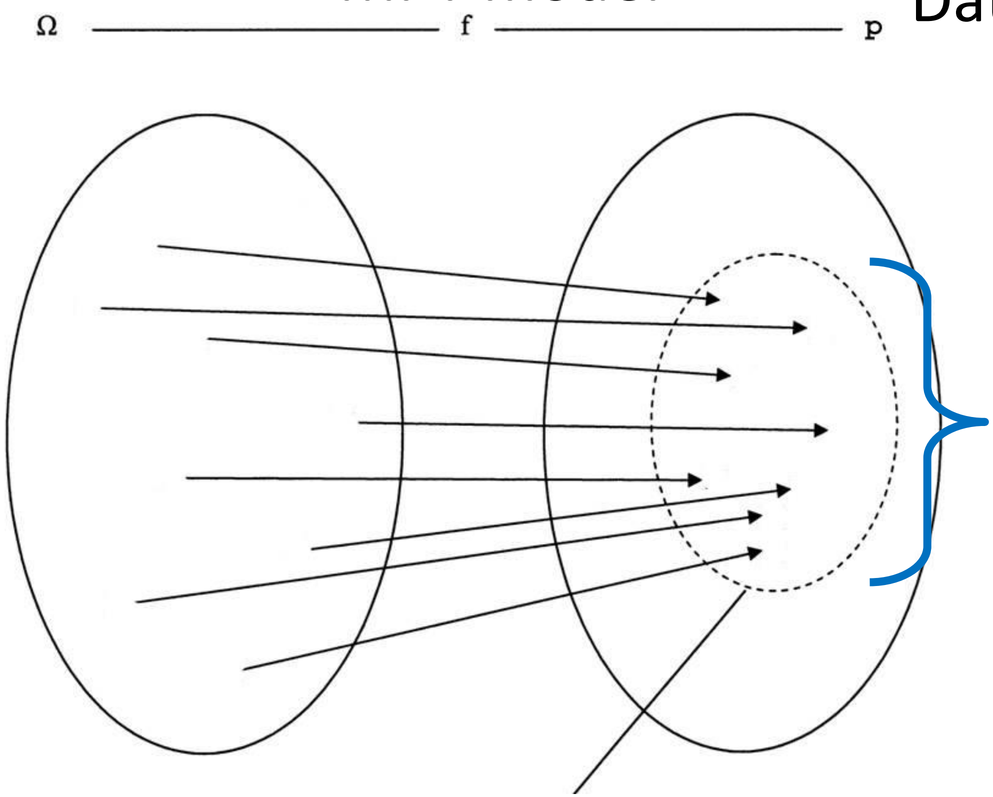
Model Assessment

- Graphical illustration of model fit:

Parameter space

MPT model

Data space



Subset of possible data compatible with the model

Model Assessment

- Goodness-of-fit test: How can we test whether a model fits the data?
- Hypothesis: the data are generated by the model
- H_0: $\pi \in f(\Omega)$

  → "the true category probabilities π are compatible with the model equations $f(\Omega)$"

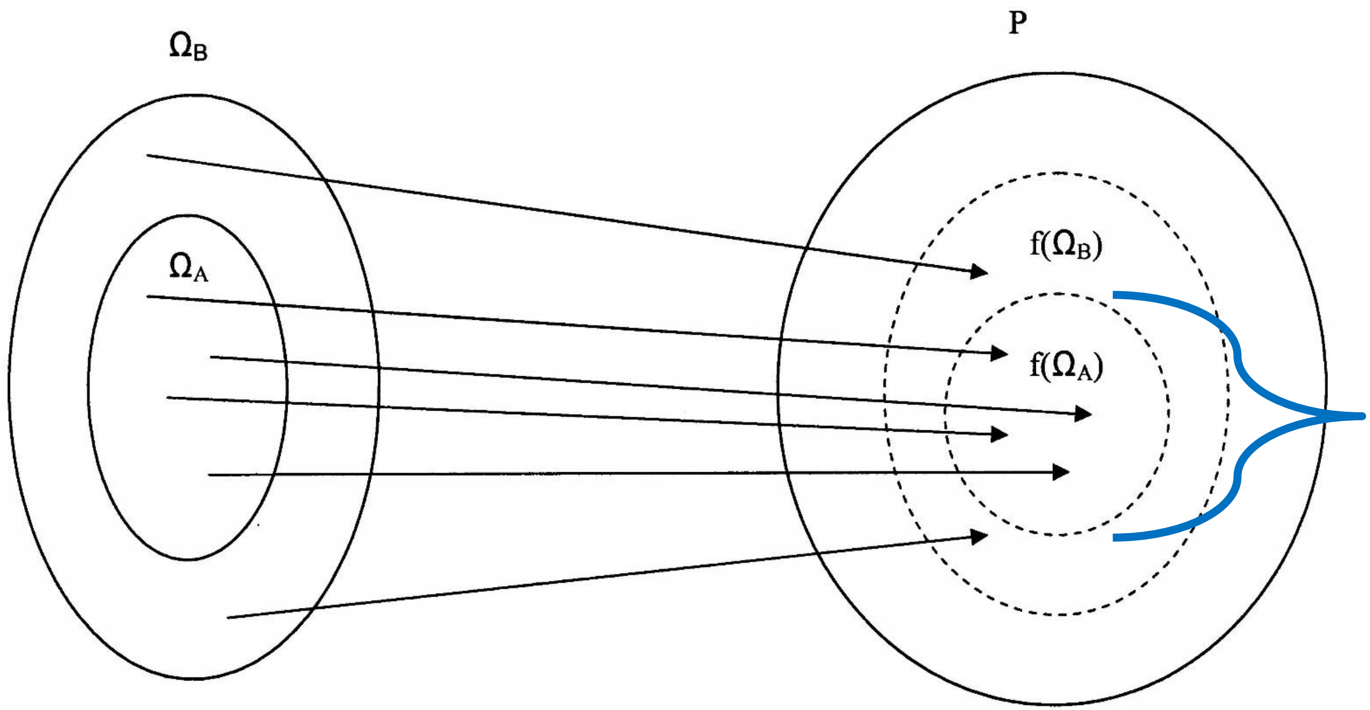
- Test statistic: Under H_0, the statistic G^2 is χ^2-distributed with degrees of freedom:

$$df = \sum_{k=1}^{K} (J_k - 1) - S$$

Model Comparisons: Nested Models

Hierarchical model families:

- Model M_A is a nested model (= special case) of M_B
- e.g., if M_A is obtained from M_B via parameter restrictions



Subset of possible data compatible with both MPT models

Model Comparisons: Hierarchical Model Families

- If model M_A is nested in M_B then
  - G^2_A is χ^2-distributed with df_A
  - G^2_B is χ^2-distributed with df_B
  - $\Delta G^2_{A\text{-}B} = G^2_A - G^2_B$ is χ^2-distributed $df_{A\text{-}B} = df_A - df_B$
- Hence, we can use $\Delta G^2_{A\text{-}B}$ to compare nested models using χ^2-tests
- Unfortunately, $\Delta G^2_{A\text{-}B}$ cannot be used for non-nested models
  - Solution: Information-theoretic measures (AIC, BIC)

Model Selection: Information-Theoretic Measures

- Core idea: Select the model that achieves the best tradeoff between model fit vs. model complexity
- Akaike Information Criterion (AIC)
  - $AIC(M_0) = -2 \log(L(\boldsymbol{\theta};\mathbf{n})) + 2\,S$
- Bayesian Information Criterion (BIC)
  - $BIC(M_0) = -\,2 \log(L(\boldsymbol{\theta};\mathbf{n})) + S \log(N)$
- Application: Choose the model with the smallest AIC/BIC

- To assess model fit: Comparison to saturated model
  - $\Delta AIC(M_0) = AIC(M_0) - AIC(\text{saturated}) = G^2(M_0) - 2\,df(M_0)$
  - $\Delta BIC(M_0) = BIC(M_0) - BIC(\text{saturated}) = G^2(M_0) - df(M_0) \log(N)$
  - Model fit is good if $\Delta AIC < 0$ or $\Delta BIC < 0$

Basics of MPT Modeling

Overview:

1. Formal model structure
2. Identifiability
3. Parameter estimation
4. Model assessment & comparison
5. Appendix: The power divergence statistic

Power Divergence Statistic

- Alternative distance measure: Pearson X^2

$$\chi^2(\theta) = \sum_{j=1}^{J} \frac{[n_j - N \cdot p_j(\theta)]^2}{N \cdot p_j(\theta)}$$

- Both types of distance measures (G^2 and Pearson X^2) are special cases of the Power-Divergence-family (PD$_\lambda$ statistics) (Read & Cressie, 1988):

$$PD_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{k=1}^{k} \sum_{i=1}^{J(k)} n_{j(k)} * \left[\left(\frac{n_{j(k)}}{e_{j(k)}} \right)^{\lambda} - 1 \right]$$

Note that:

- Pearson $X^2 = PD_{\lambda=1}$
- Likelihood-ratio statistic $G^2 = \lim_{\lambda \to 0} PD_\lambda$

What is the best goodness-of-fit statistic?

- In case of small sample sizes, $PD_{\lambda=1}$ and $PD_{\lambda=2/3}$ outperform other PD_{λ}-statistics in terms of accuracy of χ^2 approximation (cf. Read & Cressie, 1988).
  - However, small samples are typically less of a problem in MPT model applications (unless models are tested for single participants)
- Given the fact that G^2 is a by-product of ML-parameter estimation, G^2 (= $PD_{\lambda=0}$) can be recommended for moderate to large sample sizes.
  - However, G^2 cannot be applied for samples with zero cells
  - Remedies:
    a) Ignore zero cells
    b) Add constant ε to all counts